1

DOCUMENT RESUME

ED 157 944 TM 007 427

AUTHOR Curry, Allen R.; Riegel, N. Blyth
TITLE Latent Trait Theory in the Affective Domain--Applications of the Rasch Model.
PUB DATE Mar 78
NOTE 26p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Toronto, Ontario, Canada, March, 1978)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage.
DESCRIPTORS *Affective Tests; Goodness of Fit; *Item Analysis; *Mathematical Models; Measurement Techniques; Scores; *Testing Problems; Test Items
IDENTIFIERS Item Characteristic Curve Theory; *Latent Trait Models; *Rasch Model

ABSTRACT

The Rasch model of test theory is described in general terms, compared with latent trait theory, and shown to have interesting applications for the measurement of affective as well as cognitive traits. Three assumption of the Rasch model are stated to support the conclusion that calibration of the items and tests is independent of the examinee sample, and that measurement of the trait is independent of the items used, a condition called specific objectivity. Item characteristic curves are defined and related to item slope and item discrimination. The procedure for testing the fit of items to the model is discussed. The test of whether an examinee's responses fit the model may be used to indicate whether a particular score is valid. The test of item fit may be used to investigate the nature of affective variables. A calibrated item pool is useful for repeated testing and for using as a standard for new items. The Rasch model is useful for relating a series of tests to a single longitudinal scale. Thirty bibliographical references are appended. (CTM)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED157944

U S DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Latent Trait Theory in the Affective Domain -- Applications of the Rasch Model

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Allen R. Curry

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

by

Allen R. Curry and N. Blyth Riegel

Systems Oriented Measurement and Analysis

Raleigh, North Carolina

Paper presented at the annual meeting of the National Council on Measurement in Education, Toronto, March, 1978

Printed in U.S.A.

TM007 427

LATENT TRAIT THEORY IN THE AFFECTIVE
DOMAIN -- APPLICATIONS OF THE RASCH MODEL

While interest in latent trait theory is growing among psychometricians, it appears that the focus of this interest is primarily in cognitive measurement. This focus is not unreasonable, considering that latent trait theory was developed and has been presented using achievement and intelligence testing as the base. However, application of the theory is not limited to cognitive testing, in fact, shows considerable promise in meeting many affective measurement priblems. The purpose of this paper is to present one of the latent trait models, the Rasch model, from an affective measurement viewpoint and to present applications of the model in affective testing. Sections of this paper include the following tpoics; the Rasch Model, Unidimensionality, Examinee Misbehavior, Item Characteristic Curve Theory and Item Discrimination, Item Slope and Item Discrimination, Tests of Fit to the Model, and Problems and Applications. This latter section is concerned with unreliability of scores-problems of person fit, test development, referencing trait measures, nontraditional "items," and creating longitudinal tests.

The Rasch Model

Overview

In developing a structural model for test items, Rasch's objective was to focus on individuals rather than groups of persons.

> Individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated. It is further essential that comparisons between individuals become independent of which particular instruments--tests or items or other stimuli--within the class considered have been used. Symmetrically, it ought to be possible to compare stimuli belonging to the same class--"measuring the same thing"--independent of which particular individuals within a class considered were instrumental for the comparison. (1960, p. vii)

The independence of person comparison from the items and the independence of item comparisons from the persons may be termed specific objectivity, specific so that this objectivity is distinguished from scoring objectivity (c f., Bashaw, 1970). Specific objectivity is the consequence of meeting the assumptions of what Rasch calls the simple item-analysis model.

Specific objectivity indicates that the calibration of the items (determination of difficulty or affectivity) is independent of the examinee sample,and measurement of the latent trait (ability or attitude) is independent of the items used in that measurement. These two facets of specific objectivity have been termed "person-free," "sample-free," and "group-free" for item estimates that are invariant across changes in the composition of examinee groups. "Object-free," "instrument-free,: "item-free," and "test-free" are modifiers often encountered when discussing this invariance of the person parameter estimates with respect to changes in the items used for measurement (c f. C.C. Rentz, 1975).

Paraphrasing C. C. Rentz (1975), the three assumptions of the model are these:

1. When person j attempts item i there is a probability of a correct response, P_{ij}, which may be expressed as

$$P_{ij} = C_{ij}/(1 + C_{ij}), \text{ where } C_{ij} \geq 0 \quad (1)$$

2. The value for such a case (C_{ij}) is the product of two factors,

$$C_{ij} = A_j E_i \quad (2)$$

where A_j is a characteristic of the person and E_i is a characteristic of the item. These parameters may be viewed as person ability and item easiness (c f., R.R. Rentz, 1974; Wright, 1968).

3. Given the parameters, all responses are stochastically independent.

By substituting $A_j E_i$ for C_{ij} in equation (1) we obtain the probability of a correct (positive) response expressed as a person factor and an item factor

$$P_{ij} = A_j E_i/(1 + A_j E_i). \quad (3)$$

Conversely, the probability of an incorrect(negative) response (Q_{ij}) is $1-P_{ij}$; therefore,

$$Q_{ij} = 1 - (A_j E_i/(1 + A_j E_i)) = 1/(1 + A_j E_i). \quad (4)$$

The value of the product $A_j E_i$ represents the odds of success of person j on item i. Therefore, if we wish to compare two persons, we could compare their odds of success on one item, i.e., $A_1 E_i/A_2 E_i$. As noted by Rasch (1960), if we use other items for this comparison the odds for each person could change, but the ratio of the two odds would remain the same, e.g., $A_1 E_i/A_2 E_i = A_1 E_{i'}/A_2 E_{i'}$. Similar statements can be made concerning the item parameters. Such comparisons of

persons or items may be termed specifically objective or invariant. Furthermore, due to parameter separability in the model, see equation (5), it is possible to estimate person parameters separately from item parameters and vice versa. "Since the equations can be applied to any subset of persons or items, it is not only possible to estimate person parameters independent of item parameters but also independent of the parameters of other persons" (C. C. Rentz, 1975, p. 21).

One final manipulation to the probability equation is necessary in order to demonstrate the separability of parameters and to render the formulation comparable to that found in the literature and in computer programs for Rasch analysis. This manipulation is taking the natural logarithm (ln) transformation of equation (3):

$$P_{ij} = \exp(\Theta_j + e_i)/1 + \exp(\Theta_j + e_i), \qquad (5)$$

where $\Theta_j = \ln A_j$ and $e_i = \ln E_i$

Thus the product, A_jE_i, is separated into two additive components, one pertaining to person ability (attitude) and the other reflecting the degree of item easiness (affectivity).

It should be noted that inspection of equation (5) shows that in order for $P_{ij} = \frac{1}{2}$, the ability of person j should equal the negative of item easiness, $\Theta_j = -(e_i)$. That is, a sign reversal is necessary to achieve equality of the parameters. In order to alleviate this slight inconvenience, a difficulty scale will be employed in this investigation. Thus (5) becomes

$$P_{ij} = \exp(\Theta_j - d_i)/1 + \exp(\Theta_j - d_i), \qquad (6)$$

where $d_i = -(\ln E_i)$.

Some Considerations of the Model

As given above, the assumptions of the model appear to be clear-cut; however, these assumptions imply conditions which must be met in order for the model to apply (i.e., for the data to fit the model). Among these conditions are unidimensionality of the trait being measured, equality of item discriminations, and no guessing. The condition of unidimensionality is discussed by Coffman (1966), Hambleton (1969), Lord and Novick (1968), and Panchapakesan (1969); the topic is reviewed by C. C. Rentz (1975). Guessing in latent trait models is also reviewed by C. C. Rentz (1975) with viewpoints elaborated by Hambleton (1969), Lord and Novick (1968), Panchapakesan (1969), Ross (1966), and Wright (1968). The operation of item discrimination is given in detail by Lord and Novick (1968). However, an introduction to each of these topics has been included in this paper.

Unidimensionality

The number of dimensions of the latent space depends upon the number of traits (abilities, attitudes) measured by collection of items in the population of examinees of interest. A single trait dimension is required by the three most popular latent trait models -- normal ogive, Birnbaum (two parameter logistic), and Rasch models. It has been suggested that factor analytic techniques be used to determine if unidimensionality can be assumed (Hambleton & Cook, 1977; Hambleton & Traub, 1973). However, as noted by Hambleton, et al. (1977) this procedure may lead to difficulties. Furthermore, R. K. Rentz states that the best way to determine if a set of data meet the assumption of unidimensionality, at least under the Rasch model, is to examine direct tests of model-data fit; the relationship of other statistical techniques, such as factor analysis, to unidimensionality are not yet clear (1977, p. 10).

Other factors relating to the detection of unidimensionality are also discussed by R.R. Rentz.

> The (Rasch) model requires an item's easiness to be expressed along a single dimension; it also requires a person's ability to be expressed along a single dimension. Furthermore, there is a kind of symmetry between these two dimensions to the extent that, among other things, we may infer the nature of the particular "ability" of the person from what we can observe about the items [similar to the way factors are interpreted in factor analysis]. These inferences are a way of attaching semantic lables to what are essentially psychological constructs. Yet, it seems reasonable to assume that a group of arithmetic items measure an "arithmetic ability", and a group of items that require a person to answer questions about material he has read measure something that could be labelled "reading ability". A problem frequently arises when we try to merge mathmatical ideas [i.e., the parameters of the model] with our psychological and semantic ideas. The problem is that we mistakenly assume that psychological unidimensionality, semantic unidimensionality, and mathmatical unidimensionality refer to the same unidimensionality. The fact is, they may or they may not. Simply because we can assign different semantic labels to seemingly different groups of items [i.e., comprehension and vocabulary or addition, numerical reasoning, and geometry] does not guarantee that they represent either valid and different psychological constructs or mathmatical dimensions. Such issues are ultimately empirical. (1977, p. 9)

Examinee Misbehavior

Guessing, faking, and other similar "misbehaviors" are troublesome in both achievement and affective measurement. Wright (1968, 1977) suggests that we should not allow such factors to influence our measurements. Other researchers would seem to advocate changing our measurement models so as to account for these factors (e.g., the three parameter, logistic model, which accounts for variance in item discrimination and adds a so-called "guessing" parameter). While development and study of latent trait models is still too new to settle these two philosophical points, one issue is clear -- the mathematics of estimating the Rasch model parameters is presently avaiable for use. The mathematics of other models is, at best, troublesome and require to these authors what seem

to be very arbitrary assumptions (eg., setting of limits upon discrimination values -- see Lord, 1968 and 1975; Wright, 1977).

All experienced test users and developers are well aware that guessing, faking, etc. occur. Now if we employ the Rasch model, as it seems we must or forego the benefits of latent trait theory, how do we account for such problems? Methodology for such an accounting is given in a section of this report.

Item Characteristic Curve Theory and Item Discrimination

Central to the development of item-characteristic curve theory is the concept of a latent trait (c f., Lord & Novick, 1968). Latent traits or abilities are hypothetical constructs which are used to explain an individual's behavior. Given that we need only one latent trait to account for the behavior, say, test performance, the non-linear regression of the item score on the latent trait is called an item characteristic curve (ICC). If the item of concern is dichotomously scored, then the ICC specifies how the observed responses depend upon the latent trait (Lord & Novick, 1968, p. 360). In practice we know how the examinees respond to the item, and if we make certain assumptions about the shape of the ICC, then we may make inferences about the unobservable latent trait.

The shapes or mathematical forms of the ICC are specified when we assume a measurement model, e.g., normal ogive or logistic. When we adopt a model, we dictate a general shape for all ICC, but the "location" and "steepness" of each item's curve will depend upon the

specific attributes or characteristics of the item. Given a shape for the ICC and given that the item is dichotomously scored, the non-linear regression of the item on the latent trait provides the probability of success on that item for various values of the latent trait being measured.

To develop ICC under the normal ogive model, we must assume that there is only one ability, θ, needed to explain the behavior of interest and that θ can be scaled so that the regression of the dichotomously scored item on θ (i.e., the ICC) follows the normal ogive form (Lord & Novick, 1968, p. 362). The conditional probability of a correct response on item i for a subject with θ_j ability is given by

$$P_{ij}(x_{ij} = 1|\theta_j) = \Phi L_i(\theta_j) \qquad (7)$$

where $x_{ij} = 1$ denotes a correct, observed response,

Φ is the cumulative normal distribution function whose argument is a linear function of θ_j, $L_i(\theta_j)$, and

$L_i(\theta_j) = a_i(\theta_j - b_i)$.

The plot of the P_{ij} values corresponding to differing values of θ produces the expected or theoretical item characteristic curve for item i whose location is b_i. The "steepness" of this ICC is directly related to the remaining item parameter, a_i.

The ICC has a point of inflection at $\theta_j = b_i$, which corresponds to $P_{ij} = 0.5$. Thus b_i "locates' item i with respect to the ability scale. The rate of increase or slope of the curve, β_i, at this point is directly related to a_i,

$$\beta_i = a_i/\sqrt{2\pi} \qquad (8)$$

The role of a_i in the measurement process may be illustrated by examples. Consider the situation where $a_i = 0$: this results in the

probability values for all θ being .5. That is, no differences between values of θ can be detected; we are unable to discriminate between subjects of different ability. In contrast, an extremely high a_i, say 100, would provide almost perfect discrimination. Those subjects "below" the item ($\theta_j < b_i$) would almost surely miss the item ($P_{ij} = .01$), while those subjects "above" the item ($\theta_j > b_i$) would be expected to give a correct response ($P_{ij} = .99$). Thus "a_i is a parameter that indicates the quality or value of an item in the basic sense of the amount of information the item provides about θ" (Lord & Novick, 1968, p. 367).

When considering logistic models, "a_i and b_i are item parameters whose roles are generally the same as those of the item parameters in the normal ogive model" (Birnbaum, 1968, p. 400). Therefore, we may also conceptualize a_i as item discrimination or as an index of item quality under the two-parameter logistic (Birnbaum) and Rasch models. Following this conceptualization, Birnbaum investigated the condition of equality of item discriminations implied by the Rasch model. Birnbaum presents evidence from one test that items differ in their discriminations, and thus he implies that a model requiring equality would be of little practical utility (1968, p. 403).

While Birnbaum's conclusions are also reached by Hambleton (1969) and Lord (1972), other researchers have adopted differing viewpoints. Some researchers believe that the enhanced power and simplicity of the Rasch model indicate that test items should be selected to meet the discrimination condition (e.g., Angoff, 1971; Bock & Wood, 1971). Other researchers have noted that strict equality may not be necessary but that some variation in item discriminations may be tolerated by

the model (c f., Curry, 1977a; Panchepakasan, 1969; C.C. Rentz, 1975; Rentz & Bashaw, 1975; Wright & Panchepakasan, 1969).

Item Slope and Item Discrimination

A review of Rasch model research shows a common usage of the term item slope rather than item discrimination. The use of item slope stems from Rasch's research which was conducted independently of latent trait theory. In the development of his model, Rasch relied upon, not item characteristic curves, but upon the regression of "item log odds" on "test log odds." As noted earlier, the ICC is based upon the regression of item probability of success on ability. Item log odds may be obtained from the item probability by

$$\ln O_{ij} = \ln (P_{ij}/1 - P_{ij}). \qquad (9)$$

Test log odds is actually the average odds for success on all items in the test. This average is equivalent to ability. Thus one major difference between the two regressions is the change from item probability to item log odds. The other difference is that the "log odds" regression produces a straight line graph. Rasch required that these log odds plots produce points which cluster around a line of unit slope (1960, p. 172). Thus the concern among Rasch researchers with item slope values.

Why does Rasch require lines of unit slope? The answer is found in the philosophy or goals used to develop the model. Rasch desired to produce measurements so that statements such as "person A is twice as able as person B" would have meaning. This implies that under the model, for an item of given difficulty, person A's odds for success in that item would be twice person B's odds. If this relationship is "more than an ad hoc description of a limited set of data", the

relationship should be in force in a more general context (Rasch, 1960, p. 9). Therefore, if we average over items in a test (items measuring a common trait), the relationship should be maintained; person A's average odds will be two times the average odds for person B. In plotting the transformed (by logs) item odds vs transformed average odds, the proper relationship will be depicted only if the plotted points fall along a straight line. Furthermore, since the two axes share a common metric, log odds, the slope of the line must be unity.

The question now becomes one of determining the relationship between the "log odds slope" and the ICC slope and discrimination. According to Panchapakesan (1969, p. 27), in the estimation of model parameters the origin of the difficulty scale is arbitrary. She further states that this origin is commonly set (to zero) by requiring the difficulties to average to zero. Also in Rasch analysis, the average item discrimination is set to unity with this information, let's consider the average item in a test, an item with a difficulty of zero and a discrimination of unity.

The plot of the regression of item log odds on test, average, log odds extends on both axes from minus infinity to infinity. Furthermore since this regression plot must intersect a line perpendicular to the zero point on the test log odds axis (because difficulty is zero), the line connecting the plotted points will be of unit slope. Thus, the value of the log odds slope is the value of item discrimination under the logistic scale. [The interrelationship between the normal ogive, Birnbaum, and Rasch models is given by Curry (1977a) and Curry, Bashaw, and Rentz (1978).]

Tests of Fit to the Model

The Rasch model provides a theoretical or expected data matrix under the conditions that the data may be adequately represented by the model. We have at hand the actual or observed data. The difference between observed and expected is highly reminiscent of a chi square statistic. Employing this statistic, we may perform a test of fit between our data and the expected values predicted from the model (see Wright, 1977; Wright & Panchepakasan, 1969, for calculation details).

The above is how it may be determined if an entire set of data differs from the expectation. To make this investigation for an individual item, we must resort to an approximate chi square. This item chi square, transformed by dividing by its degrees of freedom, is termed mean square fit (Wright & Mead, 1976). However, since mean square fit (MSF) is only approximate, items should not be discarded on the basis of this statistic alone.

Some researchers have assumed that item slope or discrimination will reflect all failures of the item to fit the model. While this point is in need of empirical verification, discrepant slope (or discrimination) values are sufficient to indicate misfit.

R.R. Rentz (1974) proposes a bivariate plot of item slopes and MSF as in Figure 1. This plot typically produces the "kidney" shape shown with the values centered upon unity for both slope and MSF. R.R. Rentz (personal communication, 1975) makes these recommendations:

a) If you have an abundance of items and want to choose the few best fitting, select those items within the circle (Figure 1, dashed lines).

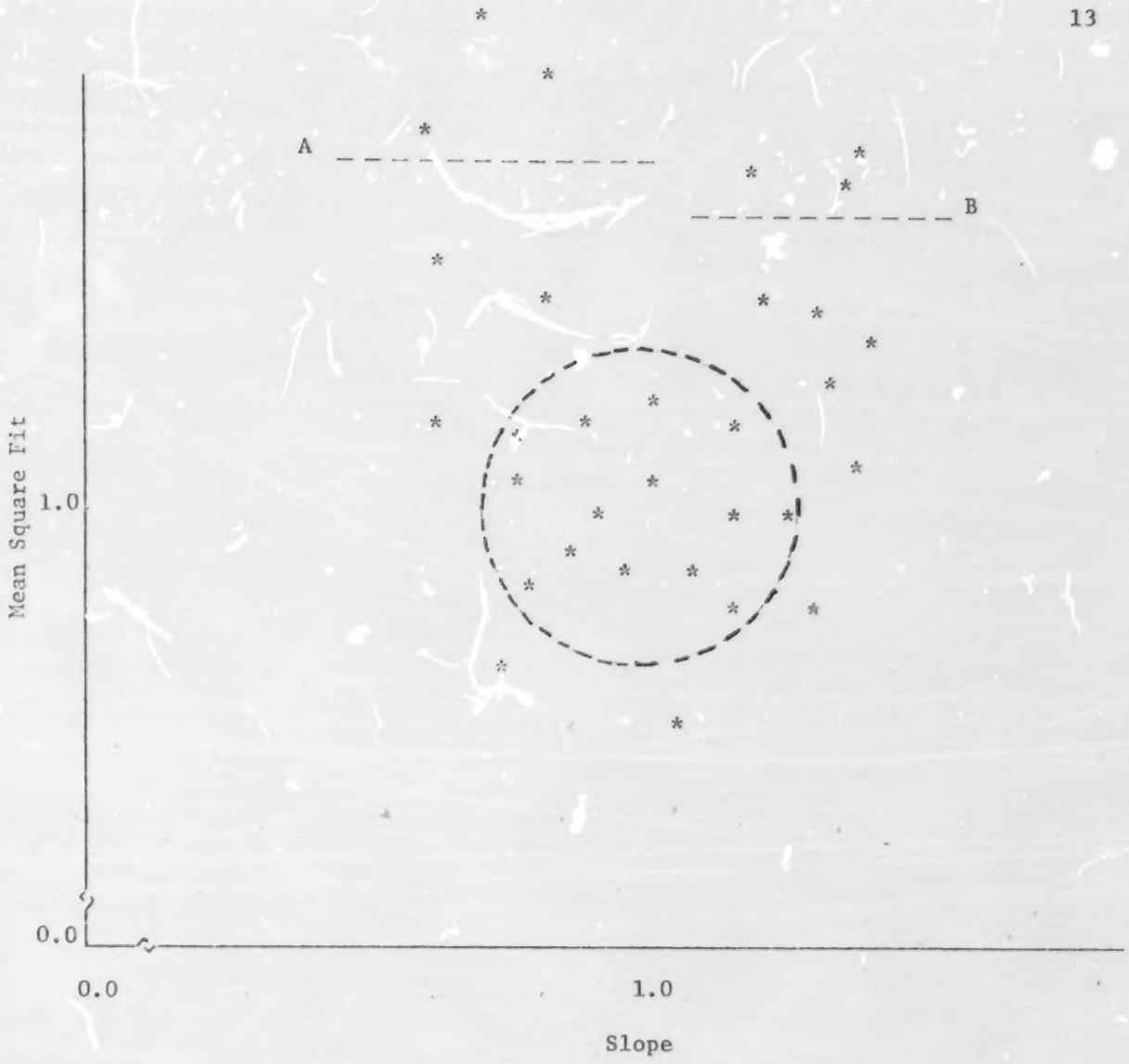

Figure 1. Example of a Plot of Item Slope and Mean Square Fit

b) If you desire to retain most of the items, discard only those that are the worst fitting, those above the lines A and B.

Experience has shown that slope values which fall between approximately .75 and 1.25 may be considered as showing fit to the model (Curry, 1977a; Panchepakesan, 1969; C.C. Rentz, 1975). These are conservative values; it may be that more misfit can be tolerated by the model. According to Wright(cited in Ryan, 1977) and C.C. Rentz (1975), mean square fit values of approximately 3.0 do not indicate too great a degree of misfit. However, this value may fluctuate as MSF is affected by examinee sample size.

An additional concern beyond how well items fit the model is the degree to which examinees fit the model. That is, do the individual examinee's responses conform to our expectations under the model or do they indicate "misbehavior"? R.R. Rentz suggests an index of person fit which may be used to tell us whether or not a score is valid or not, and whether it should be discarded (1977, p.15). Applications of this person fit (PF) index will be given later; however, the formulation is presented here for completeness.

The PF index for an examinee with j ability or attitude is given by

$$PF_j = 1/k \Sigma ((X_{ij} - P_{ij})^2)(P_{ij}(1 - P_{ij})) \qquad (10)$$

where k = the number of test items,

X_{ij} = 0/1 (incorrect/correct, negative/positive) response for the particular examinee to item i,

P_{ij} = probability of a correct or positive answer by the examinee to item i (obtainable from equation (6)),

and the summation is taken over the k items.

High values of PF would indicate high degrees of misfit; low values would indicate fit. When $P_{ij} = .5$ for all items administered to the examinee, the expected value of PF is one. Rentz further notes that while there is no knowledge as to what values are "statistically" significant, the index can be used for diagnostic information. (See Wright (1977) for another discussion of person fit.) In our opinion, a little experience with the PF index will soon show its utility as a "flag" and an indicator that an examinee's score should be examined in detail.

Problems and Applications

There are several problems inherent in affective measurement. One of the most perplexing in terms of the reliability of scores obtained is the susceptibility of responses in the affective domain to both internal and external variables. Factors such as these can alter an individual's score or profile to the extent that the attendant diagnosis would also change. Two internal variables which are particularly troublesome to the practitioner are related to role and mood differences. The "halo effect" often noted in self-report studies is a role which the subject portrays in responding to the instrument. Similarly, mood swings occurring between two testings with the same instrument, no matter how close in time to each other, can often yield 25% or more reversals in individual item responses on the total test. Besides these internal influences, external factors such as the role portrayed by the test administrator can also be potent moderating variables in affective scoring patterns.

Another, more theoretical aspect of affective measurement is the definition of the construct or trait of interest. Various theorists espouse different systems of affective variables, with unique configurations of both structure and function. Unfortunately, sometimes the con-

struct is erroneously assumed to be unidimensional and instrumentation is developed accordingly. In such cases as this, the results derived from measurement are confounded by a myriad of forces, none of which are taken into account during the initial design phase of instrumentation. Accordingly, it seems appropriate to define the behavior of the construct of interest in a theory; then to generate hypotheses which will examine specific aspects of the behavior based upon that theory; and, finally, to modify the theory and propose new hypotheses based upon the results of the administration of the test items. When the initial theory allows confounding influences to operate in the system without delineating them and trying to account for them, it is not surprising that the results of hypothesis testing efforts are inconsistent and confusing.

What is needed is a method by which the theory can be modified in an iterative sense to match reality more closely. With strong measurement techniques and rigorous experimental design, the theorists can have confidence in the hypothesis-testing results and accordingly make the necessary alterations in the theoretical framework to strengthen its predictive capacity.. After all, of what use is a theory unless it serves as a vehicle for the prediction of behavior or outcomes, given certain prior conditions and input variables?

Unreliability of Scores -- Problems in Person Fit

Untoward testing behavior, while frequently informative (for example, to suggest counseling strategies), tend to nullify measurement results. Such factors as mood swings, "halo effects," faking, and so forth, while known to bias test results, are not easily detected. However, under Rasch model theory, powerful methods exist for assessing person fit.

According to R.R. Rentz(1977), under the topic of person fit are three separate but related concepts -- test bias, test validity, and diagnosis. [These concepts were developed by Rentz in a framework of achievement testing; however, these concepts generalize to affective testing.]

For an examinee with large misfit, we may assess the validity of his or her test score. This involves determining if the collection of items (total test or part thereof) has provided a valid score on the trait to be measured. It is assumed that other examinees are validly tested.

Concerning the assessment of test bias, R.R. Rentz states:

> The model provides us with a powerful theory of what ought to happen when a person encounters an item. When our expectations fail to materialize, we have recourse to ready explanations. If certain items misbehave, we throw them out; they are biased items; they do not measure the same thing in the same way as do the other items. If certain people misbehave in the measurement situation (i.e., their misfit is large), they are not being properly measured by the test; the measurement situation is biased. The person may be getting a higher score than he deserves(perhaps by lucky guessing), or he may be getting a lower score than he would get on a more appropriate set of items. Lack of person fit is a defensible definition of test bias for an individual examinee. (1977, p. 14)

If our examinee exhibits patterns in misfit, we may use these patterns to ascertain strengths and weaknesses. As Wright(1977) notes, if for most examinees the item (or items) is "'fair',then an excessive residual (high PF index) suggests an irregularity in a person's development" (p.112). [Here Wright is referring to cognitive development but we may generalize to, say, moral-ethical, self-perception, or other interesting areas of development.] Such concerns are ones of diagnosis. "Diagnosis is nothing but the discovery of a pattern deviant from expectation, and the Rasch model provides a clear statement of that expectation" (R.R. Rentz, 1977, p. 14).

While we are still referring to problems in person fit, it would be useful to employ the concept of a calibrated item pool. Given a test

built to conform to the Rasch model or equivalently a test calibrated by and exhibiting fit to the model, we may refer to the test as a calibrated item pool. Any subset of items from the pool may be used for person measurement. Such person measurements are referenced to (on the same) total test scale, and a person's measurement is comparable to his measurement obtained from administering the total test (cf., Curry, 1977b; Wright, 1977). In short, we have all possible subsets of the items equated to a common metric. We may use this calibrated item pool to assess mood swings or to determine baseline data.

If we suspect great susceptibility to mood swings or we wish to obtain baseline data concerning reactions under a variety of situations, we could repeatedly test with only one set of items. However, we would expect must learning to take place, examinee boredom to occur, or an opportunity to "psych" out the test (faking) to present itself. To alleviate such problems, utilize a calibrated item pool and change test "forms" at will.

Test Development

Assume we have a set of items of interest (chosen by experts, empirically selected, or theoretically trustworthy), items known or believed to relate to the trait we wish to measure. We may use these items as a core of "marker items" for test development. Such development would proceed by adding more items to the marker pool (test) by the dual approach of assessing theoretical and empirical conformity. Empirical conformity could be defined as tests of fit to the Rasch model -- items which exhibit fit to the marker items, under the model, would be viewed as relating to the trait in a known amount. In such a manner, we develop a test based upon what could be well accepted measures of the construct (the marker items).

Referencing Trait Measures

Rasch model trait-level estimates may be anchored to various reference points. We may reference these estimates with respect to the test items or to a special group of people. Scaling to, say, the center of the items would result in a person with an average trait-level estimate being expected to respond positively (pass) 50% of the items in the total pool. We could, of course, center on other points -- the upper 20% of the items, lower 10%, or as we choose. Similarly, we may select a group of people as a reference point. In such a situation it could be said, " I believe Keith needs to visit me (the guidance counselor), his score indicates a high probability (obtainable from the model) that he will have self-concept problems in adolescence." Here we would have tested children with self-concept problems and use them as a "diagnostic reference point." We may use as many reference groups as necessary -- withdrawn, exhibitionist, "normal", self-actualized, etc.

Several sources explain this referencing (Curry, 1977a; Rentz & Bashaw, 1977; Whitely & Dawis, 1974; Wright, 1977). However, the point is that we can make explicit reference to a domain of items or a domain of people. Thus we are providing ready meaning to our measurements.

Nontraditional "Items"

In cognitive measurement we may calibrate into our item pool not only "regular" items (2+2=?; pronounce the word "B-A-I-T"; etc.) but also behaviors. For example,in a reading test we may wish to incorporate such behaviors as "at-grade-level in reading in the third grade" into the test. Once this is done, we can say, given a specific ability estimate obtained from only "regular" items, what is the probability of this second grade child being at-grade-level in reading next year and be able to supply an answer.

Examples of behavioral items in the affective domain are as follows: voluntarily eats lunch with minority groups (ethnic/racial bias measurements), is frequently among first chosen by peers as a player for team sports (sociability measures, cross-check on self-perception responses), and whines or cries when request not honored quickly (social maturity measurements). It may be noted that incorporating behaviors and/or reports (by peers, supervisors, family,etc.) is one method of enhancing the predictive capability of our tests. For example, assume we wish to know the probability of an employee receiving a satisfactory job rating from his or her supervisor. We would include yes/no ratings by the supervisor into the affective test and scale the total test (regular and rating items). Next given an attitude (trait-level estimate) obtained from only the regular items, the model will give us the probability of receiving a successful rating.

Creating Longitudinal Tests

This concept would be in effect if we choose to perform a vertical equating. We equate tests which span limited but different age/developmental level ranges to create a test (calibrated item pool) which spans the combined range. Such a procedure would allow tracking of children's progress, on a single trait, as development proceeds. This procedure would involve the production and verification (see Test Development) of developmental hierarchies. A demonstration of an equating in the cognitive domain is given by Rentz and Bashaw (1977); the report shows the utility and versatility of Rasch equatings.

Summary

Many strong theories may be found in a number of areas in the affective domain. It seems reasonable to match strong theories of behavior

with a strong theory of measurement -- latent trait theory. Under latent trait theory, it is hypothesized that a trait determines how examinees will respond to test items. Similarly, strong affective domain theories hypothesize determinants of psychological responses and, therefore, behavior. These strengths should compliment and reinforce each other, bolstering affective measurement and enhancing theory and its development.

REFERENCES

Angoff, W. H. Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), Educational Measurement. Washington: American Council on Education, 1971.

Bashaw, W. L. Measurement objectivity and homogeneity. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, March 1970.

Birnbaum, A. Some latent trait models and their use in inferring an examinee's ability. Part 5 of F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, Mass: Addison-Wesley, 1968.

Bock, R. D., & Wood, R. Test theory. Annual Review of Psychology, 1971, 22, 193-224.

Coffman, W. E. A factor analysis of the verbal sections of the Scholastic Aptitude Test (ETS RB 66-30). Princeton, N. J.: Educational Testing Service, 1966.

Curry, A.R. Invariance of the Rasch model ability estimates over differ- collections of items. Unpublished doctoral dissertation, University of Georgia, 1977a.

Curry, A.R. Measurement and evaluation. Unpublished manuscript, Systems Oriented Measurement and Analysis, Raleigh, North Carolina, 1977b.

Curry, A.R., W.L. Bashaw, & R.R. Rentz. Invariance of Rasch model ability estimates over differing collections of items. Paper presented at the annual meeting of the American Educational Research Association, Toronto, March, 1978.

Hambleton, R. K. An empirical investigation of the Rasch test theory model. Doctoral dissertation, University of Toronto, 1969. Available from R. K. Hambleton, University of Massachusetts, Amherst, Mass.)

Hambleton, R. K. & Cook, L. Introduction to latent trait models and their use in the analysis of educational data. Journal of Educational Measurement, 1977, 14, 75-96.

Hambleton, R. K., Swaminathan, H., Cook, L., Eignor, D. R. & Gifford, J. A. Developments in latent trait theory: A review of models, technical issues and applications. Paper presented at the annual meeting of the American Educational Research Association, New York, April 1977.

Hambleton, R. K. & Traub, R. E. Analysis of empirical data using two logistic latent trait models. British Journal of Mathematical and Statistical Psychology, 1973, 26, 195-211.

Lord, F. M. An analysis of the Verbal Scholastic Aptitude Test using Birnbaum's three-parameter logistic model. Educational and Psychological Measurement, 1968, 28, 989-1020.

Lord, F. M. Individualized testing and item characteristic curve theory (ETS RB 72-50). Princeton, N. J.: Educational Testing Service, 1972).

Lord, F.M. Evaluation with artificial data of a procedure for estimating ability and item characteristic curve parameters. Research Bulletin 75-33. Princeton, New Jersey.: Educational Testing Service, 1975.

Lord, F. M. & Novick, M. R. Statistical theories of mental test scores. Reading, Mass.: Addison-Wesley, 1968.

Panchepakesan, J. The simple logistic model and mental measurement. Doctoral dissertation, University of Chicago, 1969. (Available from University of Chicago Library.)

Rasch, G. Probabilities models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research, 1960.

Rentz, C. C. An investigation of the invariance properties of the Rasch model parameter estimates. Unpublished doctoral dissertation, University of Georgia, 1975.

Rentz, R. R. Evaluating the fit of tests to the Rasch model. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April 1974.

Rentz, R.R. Interpreting test scores and detecting achievement test bias using the Rasch model. Unpublished manuscript, Georgia State University, Atlanta, June, 1977.

Rentz, R. R., & Bashaw, W. L. Equating reading tests with the Rasch model (Report of U.S.O.E. Project No. OEC-0-72-5236, Vol. 2). Athens, Ga.: University of Georgia Educational Research Laboratory, September, 1975.

Rentz, R. R. & Bashaw, W. L. The National Reference Scale for reading: An application of the Rasch model. Journal of Educational Measurement, 1977, 14, 161-180.

Ross, J. An empirical study of a logistic test model. Psychometrika, 1966, 31, 325-340.

Ryan, J.P. Comparing classical and latent trait analyses of attitudinal data. Paper presented at the Southeastern Invitational Conference on Measurement in Education, Virginia Polytechnic Institute and State University, Blacksburg, Virginia, December, 1977.

Whitely, S. E. & Dawis, R. V. The nature of objectivity with the Rasch model. Journal of Educational Measurement, 1974, 11, 163-178.

Wright, B. D. Sample-free test calibration and person measurement. In Proceedings of the 1967 Invitational Conference in Testing Problems. Princeton, N. J.: Educational Testing Service, 1968.

Wright, B.D. Solving measurement problems with the Rasch model. Journal of Educational Measurement, 1977, 14, 97-116.

Wright, B.D., & Mead, R.J. Calfit: Sample-free calibration with a Rasch measurement model. Research Memorandum No. 18, Statistics Laboratory, Department of Education, University of Chicago, 1975.

Wright, B. D. & Panchapakesan, N. A procedure for sample-free item analysis. Educational and Psychological Measurement, 1969, 29, 23-48.